IEEE SA
STANDARDS ASSOCIATION

IEEE CertifAIEd™ – Ontological Specification for Ethical Algorithmic Bias

IEEE AI ETHICS

IEEE

Abstract: The IEEE CertifAIEd™ criteria for certification in ethical algorithmic bias are discussed in this ontological specification. Providing actionable methods to granularly assess and benchmark systems and organizations in their ethical performance is the goal of this work. Original methods of analyzing the respective drivers and inhibitors that influence the emergence of a quality of ethics, in this case to prevent harmful bias, are utilized by the certification methodology. The creation of the certification process is discussed, along with its intended implementation. An overview of the criteria schema and example criteria are also provided. This certification process has been designed to generate a tailorable and scalable system for the development of conformity assessment and certification for emergent ethical features of autonomous intelligent systems (AIS). The contents of this ontological specification are designed to be broadly applicable to a wide variety of domains and use-cases as well as providing flexibility through up to three levels of criteria, enabling a deeper and more sophisticated certification process where necessary.

Keywords: algorithmic bias, autonomous intelligent systems, ethics

The Institute of Electrical and Electronics Engineers, Inc.
3 Park Avenue, New York, NY 10016-5997, USA

TRADEMARKS AND DISCLAIMERS

IEEE believes the information in this publication is accurate as of its publication date; such information is subject to change without notice. IEEE is not responsible for any inadvertent errors.

The ideas and proposals in this specification are the respective author's views and do not represent the views of the affiliated organization.

Notice and Disclaimer of Liability Concerning the Use of IEEE SA Documents

This IEEE Standards Association ("IEEE SA") publication ("Work") is not a consensus standard document. Specifically, this document is NOT AN IEEE STANDARD. Information contained in this Work has been created by, or obtained from, sources believed to be reliable, and reviewed by members of the activity that produced this Work. IEEE and the IEEE Conformity Assessment Program (ICAP) members expressly disclaim all warranties (express, implied, and statutory) related to this Work, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; quality, accuracy, effectiveness, currency, or completeness of the Work or content within the Work. In addition, IEEE and the ICAP members disclaim any and all conditions relating to: results; and workmanlike effort. This document is supplied "AS IS" and "WITH ALL FAULTS."

Although the ICAP members who have created this Work believe that the information and guidance given in this Work serve as an enhancement to users, all persons must rely upon their own skill and judgment when making use of it. IN NO EVENT SHALL IEEE SA OR ICAP MEMBERS BE LIABLE FOR ANY ERRORS OR OMISSIONS OR DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Further, information contained in this Work may be protected by intellectual property rights held by third parties or organizations, and the use of this information may require the user to negotiate with any such rights holders in order to legally acquire the rights to do so, and such rights holders may refuse to grant such rights. Attention is also called to the possibility that implementation of any or all of this Work may require use of subject matter covered by patent rights. By publication of this Work, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. The IEEE is not responsible for identifying patent rights for which a license may be required, or for conducting inquiries into the legal validity or scope of patents claims. Users are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. No commitment to grant licenses under patent rights on a reasonable or non-discriminatory basis has been sought or received from any rights holder.

This Work is published with the understanding that IEEE and the ICAP members are supplying information through this Work, not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought. IEEE is not responsible for the statements and opinions advanced in this Work.

Participants

At the time this specification was completed, the IEEE CertifAIEd™,[1] Ethical Algorithmic Bias Expert Working Group had the following membership:

Ansgar Koene, *Chair*
Ali Hessami, *Technical Editor*

Martin Clancey	Aurelie Jacquet	Sara Spinelli
Cathy Cobey	Theordore Nowak	Gerlinde Weger
Yohki Hatada	R. Rama	Ali Hessami

The Algorithmic Bias Expert Focus Group

The work of CertifAIEd™ was largely driven by the efforts of expert focus groups, their appointed leads, and support from the Chair. The Bias Expert Focus Group (BEFG) was formed of volunteers from diverse backgrounds and experience, including legal, computer science, technological, organizational, auditing, and fiscal. However, other experts were invited to complement gaps identified in the profile of BEFG. The BEFG held 16 ideas capture workshops in developing the ethical bias schema, a graphical representation of factors that positively or negatively influence ethical accountability, which is set out in Annex A.

[1] IEEE CertifAIEd™ is a trademark owned by The Institute of Electrical and Electronics Engineers, Incorporated.

Introduction

The advent of automation during the industrial revolution brought about societal and business benefits in large-scale production, consistency, quality, and efficiencies that made commodities affordable. One key feature of most automation systems is the existence of human in the loop (HITL) at some stage providing oversight and control on critical aspects of the process or production. The development of *learning* machines that can perform specific tasks without using explicit instructions is now the foundation of autonomous intelligent systems (AIS) proliferating pervasively in all facets of industry, service provision, and governance. These machines rely on patterns and inductive or deductive inference, thereby raising the prospect of autonomous decision-making (ADM) by algorithmic learning systems (ALS), or ADM/ALS.

ADM/ALS offers the possibility of reducing and ultimately removing the human agent from the operation, control, and supervisory roles, thereby reducing costs and potential errors while processing a much larger number of transactions offering higher service levels. While this brings savings, efficiencies, and business benefits, the removal of the human agent from the control and oversight loop brings about uncertainties and concerns regarding trustworthiness, fairness, explicability, and rationale of the automated decisions.

The uncertainties and societal concerns over ethicality and trustworthiness of ADM/ALS in all walks of life, especially in high-risk environments such as transportation, healthcare, financial, and public services, pose a formidable challenge to the uptake and innovation in deployment of the AIS-based solutions. There is thus a desire to regulate the implementation of ADM/ALS in order to provide a safety net and assurance about potential risks and societal harms that may ensue.

From a broader ethical perspective, key areas of concern in development and deployment of ADM/ALS relate to accountability, transparency, ethical algorithmic bias, privacy, and responsible governance. To this end, the IEEE Standards Association (SA) has developed a suite of detailed criteria for evaluation, conformity assessment, and certification of these properties of ADM/ALS products and services through CertifAIEd™. This program is a key facet of the IEEE SA's Global Initiative and Ethically Aligned Design portfolio.

Contents

1. Overview

1.1 Scope

The IEEE ethics certification criteria developed for assurance of many ethical facets of the development and deployment of autonomous intelligent systems (AIS) constitute an extensive hierarchical suite, developed by a panel of competent experts through a model-based creative process. The criteria suite for ethical algorithmic bias comprises articulation of pertinent critical factors at three levels of hierarchy: Level 1, Level 2, and Level 3. They collectively constitute the entire ethical algorithmic bias suite for the purposes of conformity assessment and certification. This ontological specification provides insight into and specification of Level 1 ethical algorithmic bias factors to disseminate and enhance the understanding of IEEE's ethics certification criteria.

The ethics criteria suites are also developed from a general applied ethics perspective. The development strategy and deployment approach for these criteria provide an efficient and pragmatic approach for customization of a given suite for application-specific context and requirements. This is referred to as *profiling* and, in practice, the generic ethical algorithmic bias suite can be customized into many profiles appropriate to the requirements, terminology, context, and priorities of a given sector, culture, or application vertical. This specification examines the generic ethics for ethical algorithmic bias.

1.2 Purpose

This ontological specification discusses the development and specification of ethical algorithmic bias conformity assessment and certification criteria of IEEE CertifAIEd™.[1]. The criteria are applicable to all concerns relating to algorithmic bias within the context of AIS.

2. Definitions, acronyms, and abbreviations

2.1 Definitions

For the purposes of this document, the following terms and definitions apply.

algorithmic bias: Automated recommendations and predictions that disproportionately favor a stakeholder entity over another. This may be a negative *unethical* bias that prevents fair access to education, employment, health care, and economic enfranchisement. It may be a positive *ethical* bias that weights the AIS and its data use to recommend and predict fair outcomes for identified stakeholders within the context of use for the AIS.

ethical algorithmic bias: A contextual set of values pertaining to a framework of expectations that ensures algorithmic biases that negatively impact individuals, communities, and society have established

[1] IEEE CertifAIEd™ is a trademark owned by The Institute of Electrical and Electronics Engineers, Incorporated.

boundaries of acceptance to protect autonomy and freedoms, where autonomy is defined by one's capacity to direct one's life

NOTE 1—Ethical algorithmic bias may be an intentional bias that recognizes a bias requirement for there to be an AIS outcome that mitigates a harmful negative bias and preserves autonomy.

NOTE 2—Ethics is human focused, so ethical algorithmic bias is human centric/anthropomorphic.

NOTE 3—Ethical algorithmic bias overlaps with, and is largely complementary to, the aspects enforced and protected by law.

NOTE 4—There is recognition that complete avoidance of bias is not possible (e.g., all input features have some, slight correlation with protected features).

NOTE 5—The inclusion of bias considerations in the design of AIS typically results in a multiobjective optimization problem that inherently requires a balancing of ethical requirements (e.g., accuracy bias vs. bias in false-positive rates) as part of the implementation.

2.2 Acronyms and abbreviations

ADM	autonomous decision-making
AIS	autonomous intelligent system(s)
ALS	algorithmic learning system
EFR	ethical foundational requirement

3. Stakeholders

The key stakeholders of the ethical algorithmic bias of AIS are the following entities: developers, system/service integrators, system/service operators, maintainers, regulators, and the end users, that is, those impacted by the AIS (see 6.3 on duty holders).

Recognition that bias is a highly contextual property and stakeholder specific and knowledge of the affected stakeholders and the impact of the specific AIS application are required. Therefore, identifying stakeholders is only the first step towards understanding the potential bias issues that might affect them. It is necessary to consult the stakeholders directly in order to avoid unjustified assumptions regarding the priorities and sensitivities of stakeholders. Without consultation, the bias of the stakeholders influencing the AIS may increase negative bias to end users.

NOTE— An entity can be an individual, a single organization, or group of collaborating individuals and organizations. The above labels for the five groups of stakeholders are generic and can be mapped in terms of activities and influence against the life cycle but with overlapping activities. A single entity may assume multiple roles, that is, a developer may also fulfill and complete system design, integration, and maintenance.

4. Context

The IEEE CertifAIEd™ has been designed to generate a tailorable and scalable system for the development of conformity assessment and certification for emergent ethical features of AIS.

Algorithmic biases that negatively impact individuals, communities, and society are a direct infringement of our autonomy and freedoms, where *autonomy* is defined by one's capacity to direct one's life. When an AIS is negatively biased, opportunities to be all we can be, to actualize our potential, are taken away from us. Unfair, negatively biased automated recommendations and predictions prevent fair access to education, employment, health care, and economic enfranchisement. Given the proliferation of AIS across industries and integrated into our daily lives, the necessity for us to trust AIS outcomes is foundational for a fair and just society.

The CertifAIEd™ ethical algorithmic bias criteria suite comprises a holistic and systemic set of factors required in decision-making, rulemaking, enforcement, redress, operational governance, and, most importantly, human capacity and behavior across not only the AIS life cycle but with assumptions and dependencies from the wider AIS ecosystem as well. Taking the context of use within the broader sphere of the AIS ecosystem is necessary because bias is highly context specific; stakeholder and AIS impacts may be missed with narrow delineations of context.

The criteria have also sought to emphasize the importance of continuous monitoring to ensure appropriateness and timeliness of interventions. For example, changes to an AIS ecosystem may alter its outcomes and bias with respect to end users. Furthermore, for the purposes of accountability, this suite of ethical criteria reflects an effort to have responsibility remain with the humans and human organizations involved in the actions that will bring AIS into being as it is still seen as premature to preassign any such responsibilities to the AIS themselves.

5. Ethical algorithmic bias factors

In considering what goals/factors contribute to the quality of transparency—in addition to the classical identification of contributory factors—we recognized a need, supported by the adopted methodology, to map those goals/factors that would detract from it also. These are referenced as *drivers* and *inhibitors*, respectively, in the transparency schema (see Annex A). The rationale being many real-world constraints can frustrate well-meaning objectives due to issues of human resourcing, management, technological limitations, and cultural change.

5.1 Drivers of ethical algorithmic bias

The six supportive influencing factors (drivers) impacting ethical algorithmic bias are the following:

a) *Organizational governance, capability, and maturity:* This driver goal deals with the organization's capability, maturity, governance processes, and political will/good faith for ethical algorithmic bias assurance.

b) *Clarity of AIS operations:* This driver goal seeks to ascertain a clear definition and the articulation and communication of the concepts and results of operation in the intended environments for AIS products, services, or systems to the relevant stakeholders.

c) *Context alignment:* This driver goal aims to ensure that the context of the AIS is understood in relation to all affected stakeholders and unjustified bias is addressed.

d) *Justified use of protected characteristics:* This driver goal aims to ensure the inclusion of protected characteristics (and evaluation against such characteristics) is clearly documented with appropriate justification for their use. This considers that within specific concepts of operation, protected characteristics may be valid and required for a fair AIS outcome

e) *System behavior monitoring:* This driver goal monitors the AIS throughout its life cycle in order to identify bias problems as early as possible, recognizing that some bias in system behavior may only become apparent after the system is in operation (and may arise due to operational factors beyond the initial development).

f) *Maintaining bias profile:* This driver goal aims to ensure there is the organizational capability to correct emerging or detected bias during development, deployment, and operation through risk management, design changes, and compensation mechanisms.

5.2 Inhibitors of ethical algorithmic bias

The one constraining influencing factor (inhibitor) impacting ethical algorithmic bias is:

— *Lack of process transparency:* This inhibitory goal concerns the lack of an adequate degree of transparency in overall decision-making, including the selection of the appropriate data sets and the sources from which the data is drawn. This lack of transparency will hinder the ability of stakeholders to assess the level of bias in the AIS performance.

Explanation of the goals and associated requirements, requisite evidence, and scale of measurement are depicted in Annex B.

6. Ethical algorithmic bias certification criteria

6.1 Bias ethical foundational requirements (EFRs)

The ethical algorithmic bias schema, in conjunction with the bias ethical foundational requirements (EFRs), enables the auditing of organizations and their autonomous intelligent technologies for the avoidance of harmful algorithmic bias with clear criteria that can be turned into a scoring mechanism. As a model-based approach, the schema captures both negative and positive aspects (inhibitors and drivers, respectively) of ethical algorithmic bias for AIS with ease of reference. It represents an efficient means of real-time creative knowledge capture as well as operating as the foundation for development of ethical algorithmic bias requirements.

The detailed bias EFRs are depicted in Annex B.

6.2 Normative and instructive bias EFRs

The bias EFRs contain a series of expected behavioral norms and instructions on how to enact aspects of the certification, without going into specifics where not strictly necessary, in order to preserve flexibility of implementation within a bounded set of principles. In this spirit, the bias EFRs depicted in Annex B are classed into *normative* (mandatory) and *instructive* (recommended) for the purposes of conformity assessment against the suite of ethical algorithmic bias certification criteria.

6.3 Duty holders of the bias EFRs

The bias EFRs depicted in Annex B are additionally noted against the specific group of duty holders for the purposes of conformity assessment. The principal groups are as follows:

- *Developer (D):* The entity (see NOTE—Clause 3) that designs and develops a component (product) or system for a general or specific purpose/application. This could be as a result of a developer's own instigation or response to the market or a client requirement. The developer is responsible for the ethical assurance of the generic or application-specific product or system and associated supply chain.
- *(System/service) Integrator (I):* The entity that designs and assures a solution through integrating multiple components, potentially from different developers, and tests, installs, and commissions the whole system in readiness for delivery to an operator. The system delivery may take place over several stages. The integrator is usually the duty holder for total system assurance and certification, safety, security, reliability, availability, sustainability, and so forth. For this, it may rely on the certification or proof of ethics from various developers or the supply chain.
- *(System/service) Operator (O):* The entity that has a duty, competences, and capabilities to deliver a service through operating a system delivered by an integrator.
- *Maintainer (M):* The entity tasked with conducting required monitoring, preventive or reactive servicing and maintenance, and required upgrades to keep the system operational at an agreed service level. Maintainer could also be charged with abortion of maintenance and disposal of the system.
- *Regulator (R):* The entity that enforces standards and laws for the protection of life, property, or the natural habitat through imposing duties and accreditation/certification.

6.4 The levels of ethical algorithmic bias certification

In order to arrive at a fair and proportionate suite of criteria for bias certification, three levels are foreseen commensurate with the risks posed and the impact of any AIS-based product, service, or system on the end user and other key stakeholder communities' health, welfare, safety, and ethical values. The levels are:

- *Baseline, low impact (LI):* The smallest subset of bias EFRs is applicable for conformity assessment.
- *Compliant, medium impact (MI):* A larger set of bias EFRs than baseline is applicable for conformity assessment.
- *Critical, high impact (HI):* Any AIS product, service, or system that presents a likelihood of injury or harm to well-being, health, safety, security, and welfare must satisfy all ethical algorithmic bias EFRs.

The level of certification is determined through a risk-profiling exercise on the product, service, or system that takes place as the first phase of the conformity assessment activities.

6.5 Required evidence

These are the types and quantity of evidence items required to satisfy the stated requirements. A single requirement may relate to one or many items of objective evidence for evaluation of the degree to which the requirement is met (satisfaction).

6.6 Evaluation of evidence

This evaluation of evidence comprises a suitable scale of measurement and scoring of the evidence. A two-tier approach to the measurement of the evidence items is adopted as follows:

a) Top-level finding: No critical findings in the detailed normative requirements/areas requiring attention for improvement.

b) Overall score: On a 1 to 5 scale (based on aggregate of satisfying sublevel goals):

5- Excels baseline requirements

4- Sustains baseline requirements

3- Meets baseline requirements (pass mark)

2- Needs improvement

1- Does not meet requirements

A score of 3 is generally considered to be a sufficient pass mark for most cases. However, certain elements that represent a particularly strong risk or that operate in a mission-critical capacity may require a higher score to be considered sufficient.

NOTE—The scale of evaluation and the typical pass mark shall be appropriate to the criticality of the requirement and the nature of the evidence and may vary for each bias EFR.

6.7 The constraints of ethical algorithmic bias certification

The certification process cannot cover every potential eventuality. Changes in technology, culture, law, consumer standards, and practices may diminish its effectiveness or applicability to support the quality of ethical algorithmic bias. Eventually, without update, the certification may drift from contemporary realities and established best practices.

Therefore, it will be important to make regular updates and amendments to the underlying concept schema where appropriate. The IEEE CertifAIEd™ team has forecast potential technological and cultural developments for a foreseeable time horizon, thereby future proofing the criteria and certification as far as possible. This has been accomplished through discussion of technologies or practices that may be prototyped presently but are not yet in common deployment or in line with established norms and best practices.

Annex A

AIS ethical algorithmic bias schema

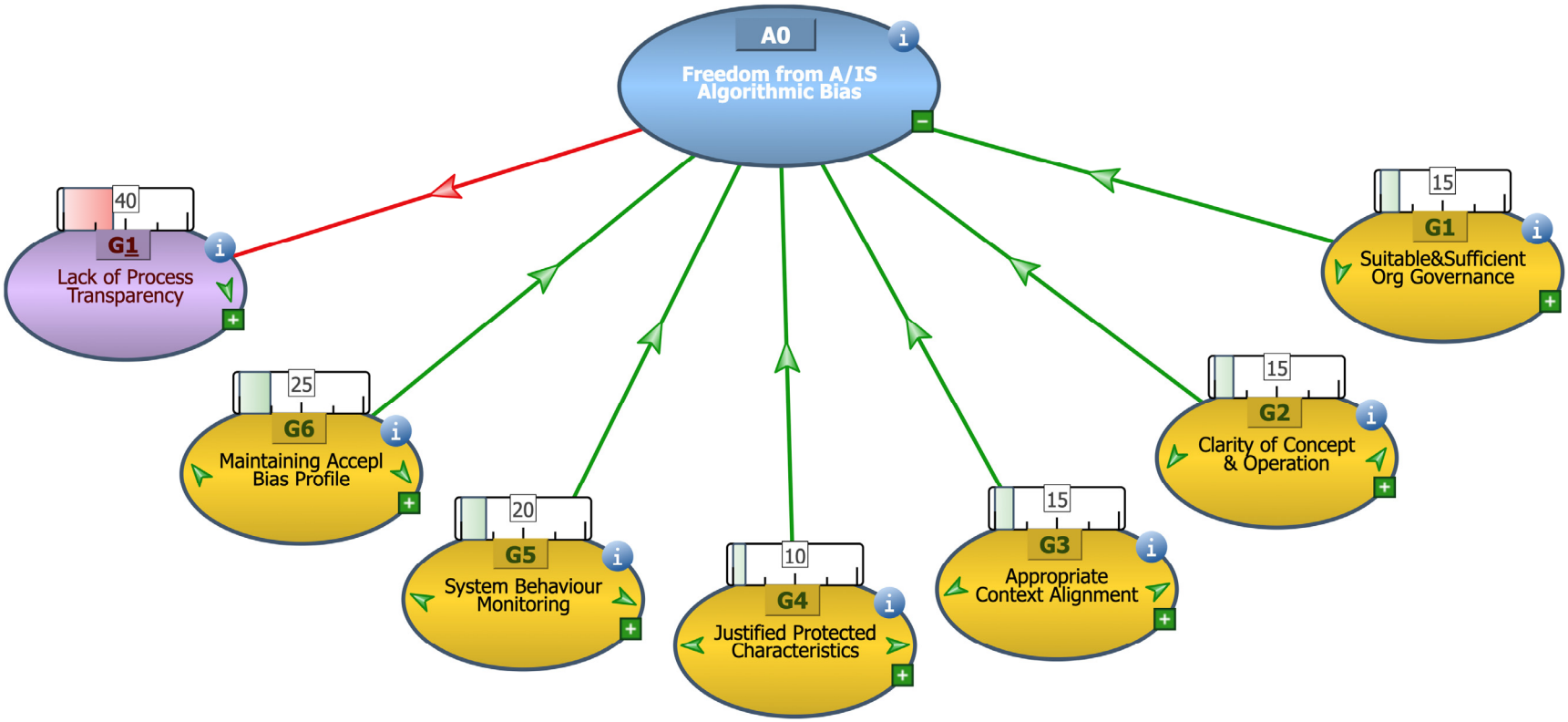

Figure A.1— Drivers and inhibitors of AIS ethical algorithmic bias.

Annex B

Ethical algorithmic bias certification criteria

Algorithmic bias schema goal description	Algorithmic bias foundation requirements (EFRs)	Normative/ instructive	Cert level LI, MI, HI	Duty holder D, I, O, M, R	Required evidence	Evidence measurement and typical pass mark
G1 - Suitable and sufficient organizational governance The organizational capability, maturity and good will/intent. Track record and reputation of the organization, its values, diversity and lack of adverse evidence showing that the organization is upholding values of fair and non-discriminatory practices.	Demonstrating the organizational good will, intent and capability/maturity to develop and deliver products and services with appropriate level of bias. Allocation of sufficient resources to address an acceptable bias in an appropriate time frame relative to the severity of the impact.	N N	HI HI	D, I, O, M, R O, M, R	The organization shall have: a) Organization chart showing lines of responsibility and accountability for maintaining acceptable bias. b) Designated positions for risk management, legal compliance, stakeholder management, and ethical bias profile management and coordination across all roles. c) Minimum assessment requirements, for each context where the AIS is used, comprising: 1. sector risks, including global operation risks (e.g., online services); 2. potential bias harms from AIS; 3. end-user needs (e.g. discrimination); and 4. supply chain awareness and compliance with minimum assessment requirements.	Two-tier approach to encourage adoption: a) Binary top-level finding: • Pass- "no critical findings in the detailed requirements". • Fail- "areas requiring attention for improvement" b) Organizational readiness finding: On 1-5 scale (based on aggregate of satisfying sublevel goals) such as: 5- Excels baseline requirements 4- Sustains baseline requirements 3- Meets baseline requirements (typical pass mark) 2- Needs improvement 1- Does not meet requirements

Algorithmic bias schema goal description	Algorithmic bias foundation requirements (EFRs)	Normative/ instructive	Cert level LI, MI, HI	Duty holder D, I, O, M, R	Required evidence	Evidence measurement and typical pass mark
					d) Implementation of local laws and requirements relevant to above minimum assessment requirements. e) Overall legal compliance (dependent on cross-jurisdictional reach and sector-specific operations of AIS). f) Engagement and participation in industry initiatives.	
G2 - Clarity of concept and operation The aims, desired outcomes, and methodological approach of the system should be clearly specified to generate a reference and highlight implicit assumptions.	The organization shall: a) Explicitly specify the purpose and application domain of the AIS system b) Define the intended user base of the AIS system c) Define an accepted performance threshold on nominal tasks d) Do due diligence in identifying potential systemic biases (positive and negative) during nominal system operation	N	HI	D, I, O, M	The organization shall provide in clear and concise terms: a) Documentation detailing the intended purpose and application domain of the AIS system. b) Documents and diagrams detailing the general methodology and pipeline followed by the system. c) Documents, test results, and audit reports supporting the acceptability and attainment of the performance threshold. d) Documents identifying potential systemic biases, where they occurred in the system, and what could help mitigate them.	Multilevel measurement on 1-5 scale: 5- Excels baseline requirements 4- Sustains baseline requirements 3- Meets baseline requirements (typical pass mark) 2- Needs improvement 1- Does not meet requirements

Algorithmic bias schema goal description	Algorithmic bias foundation requirements (EFRs)	Normative/ instructive	Cert level LI, MI, HI	Duty holder D, I, O, M, R	Required evidence	Evidence measurement and typical pass mark
G3 - Appropriate context alignment Context of the AIS is understood in relation to all affected stakeholders and unjustified bias is addressed.	The organization must not: a) Transfer systems from one context to others without alignment with the new context and without local tuning. This implies all guarantees and tests must be revisited b) Prevent/ignore new context user feedback	N	HI	D, I, O, M	The organization shall provide: a) Logs of local tuning and test results on local data. b) Evidence of feedback channels with stakeholder communities impacted by the AIS, using local languages.	Multilevel measurement such as: 2- Conformance 1- Partial conformance 0- Nonconformance
G4 - Justified protected characteristics Protected characteristics (e.g. race, sex, etc.) that are generally legally prohibited to be used as basis for discriminating between groups may under certain circumstances be valid factors to include in an algorithmic system's decision process. Medical applications, for instance, may require different procedures for male or female patients. The inclusion of information regarding protected characteristics may also be important in order to mitigate against unintended discrimination due to other factors that are correlated with protected characteristics. Therefore, the inclusion of	The organization shall provide: a) Clear identification of the types of legally/justifiably protected/sensitive characteristics that are used by the AIS b) Clarification about the purpose for which the protected characteristics are used and why it is deemed appropriate in the context for which the AIS is meant to be used c) Evidence that alternatives were explored, and an explanation of why the use of protected characteristics was determined to be the most appropriate way to	N	HI	D, I, O, M	The organization shall provide in clear and concise terms: a) Documentation enumerating the types of protected characteristics that are used. b) The purposes for which they are used, and the justification and the reasoning process for why it is appropriate and proportionate to use the protected characteristics for this purpose. c) Explanation of the anticipated consequences if the protected characteristics were not applied.	Multilevel measurement on 1-5 scale such as: 5- Excels baseline requirements 4- Sustains baseline requirements 3- Meets baseline requirements (typical pass mark) 2- Needs improvement 1- Does not meet requirements

Algorithmic bias schema goal description	Algorithmic bias foundation requirements (EFRs)	Normative/ instructive	Cert level LI, MI, HI	Duty holder D, I, O, M, R	Required evidence	Evidence measurement and typical pass mark
protected characteristics (and evaluation against such characteristics) must be clearly documented with appropriate justification for their use. This documentation must be accessible for appropriate parties (e.g., regulators and affected citizens).	proceed with the AIS					
G5 - System behavior monitoring Full life-cycle monitoring of the design, development, testing, deployment, and ongoing operation of the algorithmic system in order to identify bias problems as early as possible, with awareness that some bias in system behavior may only become apparent after the system is in operation (may arise due to operational factors beyond the initial development). Most algorithmic systems will undergo training and continual optimization throughout their service life, resulting in new or unexpected behaviors; hence, the need for behavior monitoring for ethical properties.	The organization shall: a) Have a monitoring process in place to track AIS behavior patterns to identify bias in the system outcomes as they develop b) Have an intervention plan in place for when AIS system behavior becomes unacceptably biased, including: specified intervention triggers; a protocol for how to initiate a corrective intervention c) The time frame for monitoring shall be appropriate for a system and the context	N	HI	D, I, O, M	The organization shall provide in clear and concise terms: a) Assessment of monitoring and intervention methodology (e.g., simulation results). b) Documented specification of intervention triggers with clear justification for the chosen bias thresholds. c) Evidence regarding how the monitoring/ intervention will be implemented when the system is deployed. d) Contract clauses and assigned responsibility specifying who will be held accountable in case of failure to fulfil the monitoring duties. e) Process or log for monitoring and capturing user complaints and comments f) Evidence of successful tests against benchmarks	Multilevel measurement on 1-5 scale such as: 5- Excels baseline requirements 4- Sustains baseline requirements 3- Meets baseline requirements (typical pass mark) 2- Needs improvement 1- Does not meet requirements

Algorithmic bias schema goal description	Algorithmic bias foundation requirements (EFRs)	Normative/ instructive	Cert level LI, MI, HI	Duty holder D, I, O, M, R	Required evidence	Evidence measurement and typical pass mark
G6 - Maintaining acceptable bias profile The organizational capability to correct emerging or detected bias during development, deployment, and operation through risk management, design changes, and compensation mechanisms.	The organization shall: a) Ensure that staff assigned with responsibility for the AIS must have the appropriate skills and training to be able to identify and respond to bias in the AIS b) There must be a clear process in place for notification about bias problems with the AIS and for responding to those notifications to mitigate the identified problems c) Bias considerations must be an integral part of the risk and impact assessments performed for the AIS d) Have an intervention plan in place for when AIS system behavior becomes unacceptably biased, including: specified intervention triggers based on instances or durations; a protocol for how to do a corrective intervention e) Allocate sufficient resources to address an unacceptable bias in an appropriate time frame relative to the severity of the impact.	N	HI	D, I, O, M	The organization shall provide in clear and concise terms: a) Documentation regarding the process for intervention to maintain AIS bias profile once the system is deployed, including contract clauses and assigned responsibility specifying who will be held accountable in case of failure to fulfil the intervention duties. b) Evidence of competence of the staff allocated to AIS bias profile maintenance tasks and adequacy of resources provisions. c) Records of performance before and after interventions to demonstrate effectiveness of the intervention. d) Evidence that the same bias profile is successfully maintained for each context of application.	Multilevel measurement on 1-5 scale such as: 5- Excels baseline requirements 4- Sustains baseline requirements 3- Meets baseline requirements (typical pass mark) 2- Needs improvement 1- Does not meet requirements

Algorithmic bias schema goal description	Algorithmic bias foundation requirements (EFRs)	Normative/ instructive	Cert level LI, MI, HI	Duty holder D, I, O, M, R	Required evidence	Evidence measurement and typical pass mark
	f) Ensure that the system sustains contextually					
G1b- Lack of process transparency In order to ethically deal with AIS bias issues, it is necessary to provide transparency in specifying objectives, goals and decision making, appropriate data sets, sources of data sets, clarifying/justifying the selections, and predefining the types of categorization systems and optimization metrics. The lack of an adequate degree of transparency in overall decision-making—including the selection process for the appropriate data sets and the sources from which the data is drawn—will hinder the ability of stakeholders to assess the level of bias in the AIS performance.	The organization shall refrain from: a) Obscuring key process elements that might impact on the bias profile and AIS behavior b) Using intellectual property (IP) protection as argument for blocking investigations into AIS bias c) Blocking access requests or obscuring any decisions by using secrecy, IP protection, or sensitivity as a justification Responses to access requests shall be explicit and understandable for the requesting party	N	HI	D, I, O, M	The organization shall provide in clear and concise terms: a) Accessible records of key process elements that might impact AIS bias. b) Provide evidence of having a procedure in place for explicit good-faith engagement with access requests. c) Provide audit reports showing transparency of AIS bias performance indicators.	Multilevel measurement on 1-5 scale such as: 5- Excels baseline requirements 4- Sustains baseline requirements 3- Meets baseline requirements (typical pass mark) 2- Needs improvement 1- Does not meet requirements

Annex C

Bibliography

The following sources and public domain frameworks have been consulted for the verification, coverage, integrity, quality, and currency of the certification criteria independently developed in CertifAIEd™:

[B1] *The Age of Digital Interdependence*, Report of the UN Secretary-General's High-level Panel on Digital Cooperation, United Nations, Jun. 2019.[1]

[B2] ANSI/ANS 10.3-1995, Documentation of Computer Software.[2]

[B3] "Ethics Guidelines for Trustworthy AI," High-Level Expert Group on Artificial Intelligence (AI HLEG), European Commission, Apr. 2019.[3]

[B4] BSI PAS 440: 2020, Responsible innovation – Guide.[4]

[B5] "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Apr. 4, 2019.[5]

[B6] "G20 AI Principles," *G20 Ministerial Statement on Trade and Digital Economy*, Annex, Jun. 2019.[6]

[B7] IEC/IEEE 82079-1:2019, Preparation of information for use (instructions for use) of products — Part 1: Principles and general requirements.[7]

[B8] ISO 9001:2008, Quality management systems-Requirements.

[B9] ISO/IEC 26514: 2008, Systems and software engineering - Requirements for designers and developers of user documentation.

[B10] ISO/IEC 90003:2014, Software engineering — Guidelines for the application of ISO 9001:2008 to computer software.

[B11] ISO/IEC/IEEE 29148-2011, International Standard, Systems and software engineering — Life cycle processes — Requirements engineering.

[B12] OECD/LEGAL/0449, *Recommendation of the Council on Artificial Intelligence*, May 21, 2019.[8]

[B13] "OECD Due Diligence Guidance for Responsible Business Conduct," OECD, 2018.

[B14] "OECD Guidelines for Multinational Enterprises," Update 2011, OECD, 2011.

[B15] "Sustainable Development Goals," *Transforming our World: the 2030 Agenda for Sustainable Development*, United Nations, 2015.

[B16] "Use Case Trials-Categorization of Customer e-mails Bias Assessment Report," IEEE ECPAIS, Issue 1 Draft 1, Sept. 2019.

[1] United Nations publications are available from the United Nations website (https://www.un.org).
[2] ANSI/ANS publications available from the American National Standards Institute website (https://webstore.ansi.org/Standards/ANSI/ANSIANS101995).
[3] European Commission publications are available from the Futurium website (https://futurium.ec.europa.eu/en).
[4] British Standards documents are available from the British Standards Institution website (https://standardsdevelopment.bsigroup.com).
[5] IEEE publications are available from the Institute of Electrical and Electronics Engineers, 445 Hoes Lane, Piscataway, NJ 08854-4141, USA (http://standards.ieee.org).
[6] Available from https://www.mofa.go.jp/files/000486596.pdf.
[7] IEC and ISO standards publications available from the International Organization for Standardization website (https://www.iso.org/home.html).
[8] Organisation for Economic Co-operation and Development publications available from the OECD website (https://www.oecd.org/).

[B17] “Use Case Trials-Sentiment Analysis Text Analytic Tool Bias Assessment Report,” IEEE ECPAIS, Issue 1 Draft 1, Sept. 2019.

IEEE SA
STANDARDS ASSOCIATION

IEEE CertifAIEd™

http://engagestandards.ieee.org/ieeecertifaied.html

Connect with us on:

Twitter: twitter.com/ieeesa
Facebook: facebook.com/ieeesa
LinkedIn: linkedin.com/groups/1791118
Beyond Standards blog: beyondstandards.ieee.org
YouTube: youtube.com/ieeesa

standards.ieee.org
Phone: +1 732 981 0060

IEEE

A CHAMPION FOR
IEEE AI Ethics